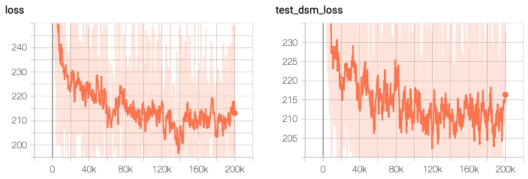
We thank all the reviewers for providing valuable feedback. In what follows, we address specific questions.

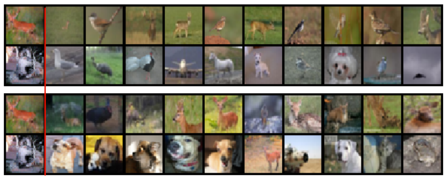
Q1 (R2): *Motivation for parameterizing the score function explicitly, rather than as the gradient of an energy model.*

The main motivation is computational. When using the gradient of an energy model as the score network, we first need one backpropagation to compute the score, and then another backpropagation to optimize the parameters. This requires higher-order gradients, which has no or limited support in many deep learning frameworks (*e.g.*, mxnet). Also, this makes computation 4 to 9 times slower due to double backpropagation (*cf.*, this Github issue), compared to directly parameterizing the score with a similar architecture. We will discuss this motivation in Section 2.1.

Q2 (R2): *Metrics or experiments to assess whether the model is overfitting or memorizing the dataset.*

Our experiment on image inpainting (Figure 5) already shows that the model is not memorizing, since we are able to generate diverse reconstructions that are different from the original unoccluded image from the datatset. As suggested by R2, we will include training/test learning curves and nearest neighbors in the appendix. For example, we provide a subset of images in the following, where the left shows the curves on CelebA, and the right shows the top 10 nearest neighbors of two samples (to the left of the red line) in pixel space (top two rows) and feature space of an Inception V3 network (bottom two rows). As expected, our model is not overfitting or memorizing the dataset.





Q3 (R2): *Issues on CIFAR-10 inception scores.*

When computing inception scores, we did not flip images. Flipping was only done in training, which is a common data augmentation technique used in training other generative models as well (*e.g.*, i-ResNet). We will clarify this more in Appendix C.2. The numbers in Table 1 are mainly from Figure 5 of the OpenAI EBM paper (arxiv: `1903.08689v1`). We agree with R2 that the inception score of WGAN-GP should be 7.86, and will correct this in the paper. We double checked other numbers in the table, and they matched numbers reported in previous work. Our inception score was computed using the original code from OpenAI, and FID score was computed with the original code of TTUR authors.

Q4 (R2, R3): *More discussion on related work.*

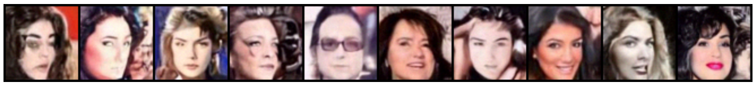
Since NeurIPS allows an additional page for camera ready, we will include a more comprehensive related work section if our paper is accepted. We will discuss Deep Energy Estimator Networks (suggsted by R3), and move the related work part in the conclusion section to this section (suggested by R2).

Q5 (R3): *Why not use continuously varying noise levels? Writing suggestions on line 33, 115, 237, 243, and 265.*

The discrete strategy is a design choice to simplify the implementation. This makes it easier for us to borrow architectural designs from existing models (we are aware of more models conditioned on discrete labels than continuous ones). We appreciate all the writing suggestions. They are very helpful and we will incorporate them in the paper.

Q6 (R3): *Scalability to higher resolution images.*

We tried modeling 64×64 CelebA images. We empirically found that our models can be scaled to higher resolution images, and the cost of sampling is the same to 32×32 images, in terms of using the same number of iterations. We will incorporate these results into the paper. Some uncurated samples are provided below.



Q7 (R3): *Extension of score-based generative modeling to discrete data.*

There are many extensions of score matching to discrete data, *e.g.*, ratio matching and minimum probability flows. We could possibly couple them with appropriate MCMC sampling methods and annealing strategies.

Q8 (R4): *Why learning with score-matching can avoid the blurriness of samples in the presence of Gaussian noise.*

During sampling, we decrease the variance of the Gaussian noise. The final variance is 0.0001, and a Gaussian perturbation with this variance is almost indistinguishable (to human eyes) when the pixel values are within $[0, 1]$. Therefore, using Gaussian noise in our model does not necessarily mean generated images should be blurry.